MÔ HÌNH TÍCH HỢP DỰA TRÊN TIẾP CẬN CƠ SỞ DỮ LIỆU LIÊN BANG, ỨNG DỤNG VÀO XÂY DỰNG HỆ THỐNG THÔNG TIN TÍCH HỢP Ở TRƯỜNG ĐẠI HỌC ĐỒNG THÁP

• ThS. Nguyễn Hữu Duyệt(*)

Tóm tắt

Vấn đề tích hợp hệ thống thông tin đã được đặt ra từ khá lâu do trong thực tế thường tồn tại nhiều phần mềm được phát triển độc lập ở những thời điểm khác nhau bởi những công ty khác nhau. Do đó, nhu cầu tổ hợp các phần mềm với nhau để có thể truy vấn được thông tin tổng hợp phục vụ cho hoạt động quản lý, điều hành của tổ chức. Bài báo này giới thiệu tổng quan về cách tiếp cận về tích hợp dữ liệu, khái niệm cở sở dữ liệu liên bang (FDB-Federated Database) và mô hình tích hợp dữ liệu dựa trên FDB. Đặc biệt, bài báo đề xuất mô hình tích hợp dữ liệu được cài đặt ở Trường Đại học Đồng Tháp.

Từ khóa: cơ sở dữ liệu liên bang, đa cơ sở dữ liệu, tích hợp dữ liệu.

1. Đặt vấn đề

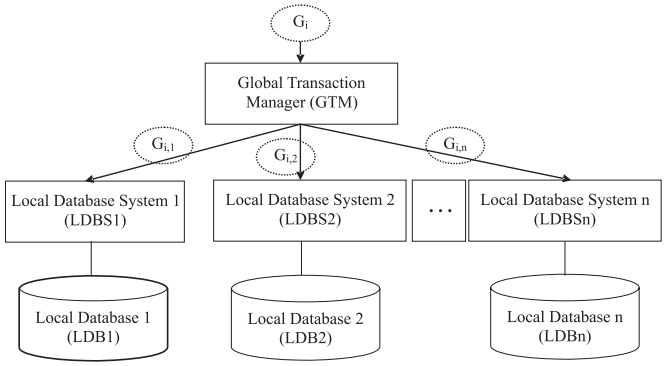
Trong thực tế, một tổ chức thường tồn tại nhiều phần mềm tương ứng với nhiều cơ sở dữ liệu (CSDL) khác nhau, phục vụ cho những chức năng khác nhau. Những hệ CSDL này thường hoạt động một cách độc lập và thường có nền tảng kiến trúc khác nhau như mô hình dữ liệu khác nhau, ngôn ngữ lập trình khác nhau... Điều này dẫn đến những khó khăn trong việc triển khai một hệ thống thông tin phục vụ hiệu quả cho công tác quản trị và điều hành hoạt động của tổ chức. Yêu cầu truy vấn thông tin từ nhiều nguồn dữ liệu khác nhau để có được thông tin có tính tổng hợp cao là một nhu cầu tự nhiên đối với bất kỳ một tổ chức nào.

Vấn đề tích hợp các hệ CSDL được đặt ra với mục đích vừa đảm bảo sự hoạt động độc lập của các hệ CSDL hiện có, đồng thời có thể liên kết giữa các hệ CSDL đó để thực hiện các truy vấn thông tin trên nhiều CSDL khác nhau.

2. Một số khái niệm liên quan đến tích hợp dữ liệu

2.1. Hệ đa CSDL

Theo [4], tích hợp CSDL là sự tổ hợp các CSDL tham gia để tạo nên một CSDL có khả năng phối hợp giữa các CSDL thành một hệ thống nhất, gọi là đa CSDL. Một đa CSDL như vậy có khả năng cung cấp các giao diện truy xuất người dùng đồng nhất tới các hệ CSDL phân tán, hỗn hợp. Hệ đa CSDL sẽ tổ hợp các hệ CSDL phân tán, hỗn hợp thành phần thành một hệ CSDL toàn cục. Trong hệ đa CSDL, các truy vấn toàn cục được phân tách thành các truy vấn con cục bộ được thi hành trên mỗi hệ CSDL thành phần. Mô hình kiến trúc cho hệ đa CSDL được mô tả như Hình 1.



Hình 1. Kiến trúc đa CSDL mức quan niệm [4]

2.2. Tính trong suốt của hệ đa CSDL

Để đảm bảo tính trong suốt trong hệ đa CSDL cần phải phân biệt giữa mô tả ngữ nghĩa dữ liệu ở mức cao hơn với việc cài đặt dữ liệu trong máy tính. Tamer Ozsu [4] đã mô tả các đặc trưng trong suốt trong hệ đa CSDL:

- Trong suốt mạng (Network Transparency): Bao gồm trong suốt đối với vị trí của các dịch vụ và dữ liệu (nghĩa là ứng dụng không cần quan tâm đến nơi cung cấp các dịch vụ và dữ liệu cho nó) và trong suốt đối với việc đặt tên (tên là duy nhất đối với mỗi đối tượng của hệ thống).

- Trong suốt nhân bản (Replication Transparency): Dữ liệu có thể được sao lặp ở nhiều nơi trong hệ thống để tăng độ tin cậy và hiệu năng của hệ thống. Một ứng dụng cụ thể không cần biết dữ liệu được truy cập từ đâu.

- Trong suốt với người dùng (User Transparency): Người dùng có thể truy cập vào hệ thống một cách đồng thời mà không quan tâm đến sự tranh chấp, đụng độ dữ liệu.

- Trong suốt hệ thống (System Transparency): Người dùng không cần quan tâm đến cấu hình máy tính và nền tảng phần mềm.

(*) Phòng Đào tạo, Trường Đại học Đồng Tháp.

- Trong suốt ngữ nghĩa dữ liệu (Data Semantics Transparency): Các xung đột về dữ liệu tiềm ẩn sẽ được giảm thiểu tối đa hoàn toàn tự động trong hệ đa CSDL.

2.3. Hệ CSDL liên bang

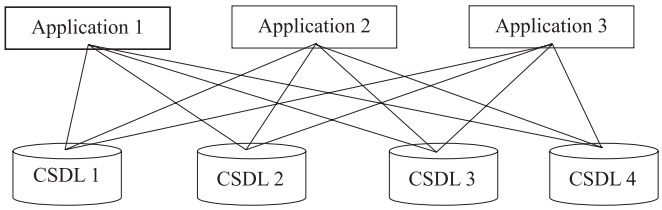
2.3.1. Khái niệm CSDL liên bang

Hệ CSDL liên bang (FDBS) là một kiểu hệ đa CSDL, ánh xạ trong suốt từ nhiều hệ CSDL tự chủ thành một FDB duy nhất. Các CSDL thành phần được kết nối với nhau thông qua mạng máy tính và có thể phân tán về mặt địa lý. Một FDB, được xem như là CSDL ảo, là một sự tổ hợp của các CSDL thành phần trong một hệ CSDL liên bang (FDBS), nhưng không có sự tích hợp dữ liệu thực sự nào trong các CSDL thành phần trong quá trình liên bang hóa dữ liệu. Thông qua sự trừu tượng dữ liệu, FDBS cung cấp một giao diện người dùng thống nhất để thực hiện lưu trữ, tìm kiếm dữ liệu từ nhiều CSDL khác nhau bằng chỉ một câu lệnh truy vấn, thậm chí trong điều kiện các CSDL hợp thành là hỗn tạp. FDBS có khả năng phân giải câu lệnh truy vấn thành các truy vấn con phù hợp với mỗi CSDL thành phần, sau đó hệ thống sẽ tổ hợp tập kết quả từ các truy vấn con. Vì các hệ quản trị CSDL (DBMS) khác nhau có ngôn ngữ truy vấn khác nhau, nên FDBS sử dụng các bộ xử lý Wrapper để dịch các truy vấn con thành ngôn ngữ phù hợp.

2.3.2. Lược đồ tích hợp dữ liệu dựa trên tiếp cận CSDL liên bang

a. Vấn đế tích hợp dữ liệu dựa trên tiếp cận CSDL liên bang

Trong bài báo này, chúng ta sẽ thảo luận chi tiết về tích hợp các hệ thống thông tin ở mức dữ liệu hay nói cách khác là tích hợp CSDL. Chúng ta sẽ tìm hiểu các tiếp cận cho phép truy cập vào nhiều DB khác nhau từ trong một ứng dụng. Cần phân biệt giữa thuật ngữ tích hợp CSDL với truy cập nhiều CSDL khác nhau. Tích hợp CSDL là hệ quản trị CSDL cung cấp chức năng xử lý dữ liệu, còn DB được tích hợp là như thể nó được quản lý bởi một hệ quản trị DB chuẩn, làm cho việc xử lý dữ liệu được hiệu quả hơn trên các CSDL thành phần. Một ưu điểm lớn hơn là hệ quản trị CSDL tích hợp hỗ trợ người phát triển ứng dụng bằng cách che dấu các mô hình dữ liệu khác nhau, các phương thức truy cập, các lược đồ. Nói cách khác, hệ quản trị CSDL tích hợp cung cấp một lớp trung gian để che dấu sự hỗn tạp, phân tán và sự tự chủ.



Hình 2. Sự chia sẻ giữa các ứng dụng và các CSDL

Các CSDL lớn thường được sử dụng bởi nhiều ứng dụng khác nhau và một ứng dụng có thể truy cập tới nhiều CSDL khác nhau. Điều này, dẫn tới một số vấn đề:

- Tất cả các ứng dụng cần biết phương thức truy cập tới các CSDL liên quan.
- Sự không đồng nhất và dư thừa dữ liệu có thể xảy ra do mỗi ứng dụng truy cập các CSDL dựa trên quy tắc ràng buộc và phụ thuộc dữ liệu mà nó quy định.
- Việc truy cập đồng thời tới cùng một dữ liệu dẫn đến sự xung đột dữ liệu nếu có nhiều ứng dụng sử dụng cùng một CSDL ở cùng một thời điểm.
- Các thay đổi trong dữ liệu và trong các ứng dụng trở nên khó quản lý.

b. Mô hình CSDL liên bang

Vì hệ quản trị CSDL liên bang là tích hợp của nhiều nguồn dữ liệu cục bộ có nền tảng kiến trúc khác nhau do chúng được phát triển độc lập dựa trên các mô hình dữ liệu khác nhau như mô hình quan hệ, mô hình hướng đối tượng hoặc XML. Nên sử dụng mô hình dữ liệu nào để biểu diễn cho CSDL tích hợp? Chúng ta khảo sát qua một số mô hình dữ liệu hiện đang sử dụng để mô tả dữ liệu.

Mô hình quan hệ: Đang được sử dụng phổ biến nhất cho các hệ quản trị CSDL tập trung và phân tán. Nhược điểm của mô hình này là khó biểu diễn cho các đối tượng dữ liệu phức tạp. Nếu sử dụng mô hình quan hệ thì sẽ khó biểu diễn các ràng buộc xảy ra trong quá trình tích hợp từ các nguồn dữ liệu khác nhau, đặc biệt các nguồn dữ liệu phi quan hệ.

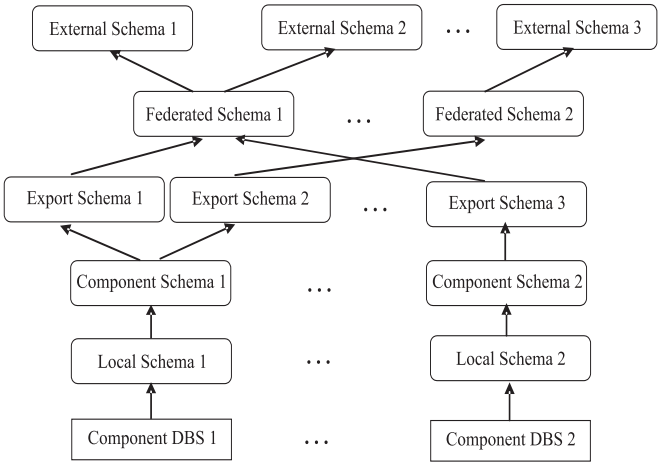
Mô hình dữ liệu hướng đối tượng: Có thể biểu diễn cho các mô hình phức tạp. Tuy nhiên mô hình hướng đối tượng chưa được sử dụng phổ biến trong thực tế vì hiện tại có rất ít các hệ quản trị CSDL hướng đối tượng có tính thương mại cao.

XML: Đang trở thành mô hình dữ liệu chuẩn được sử dụng phổ biến. XML có thể bao hàm các

phần tử chính của mô hình dữ liệu hướng đối tượng, có thể được dùng cho cả biểu diễn và truyền dữ liệu [3].

Từ những đánh giá trên có thể nhận thấy XML là mô hình thích hợp nhất để biểu diễn cho lược đồ tích hợp dữ liệu.

Kiến trúc lược đồ FDBS 5 mức [4]:



Hình 3. Kiến trúc lược đồ FDBS 5 mức

- Lược đồ cục bộ (Local Schema) là lược đồ quan niệm của một hệ CSDL thành phần.

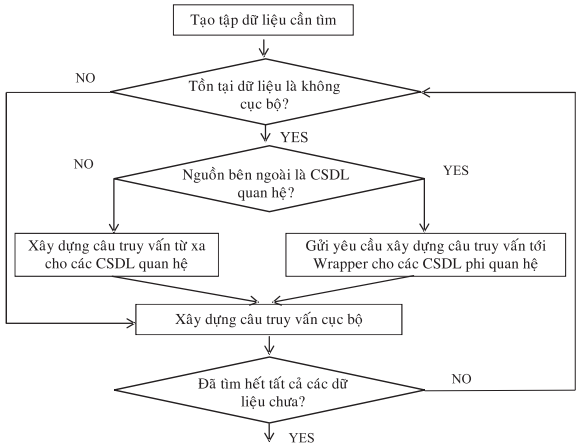
- Lược đồ thành phần (Component Schema) là lược đồ dẫn xuất bằng cách dịch từ lược đồ cục bộ thành mô hình dữ liệu tương ứng với mô hình dữ liệu chuẩn của FDBS.

- Lược đồ xuất (Export Schema) là tập con của lược đồ thành phần phục vụ cho FDBS.

- Lược đồ liên bang (Federated Schema) là sự tích hợp của nhiều lược đồ xuất. Thông tin về sự phân bố dữ liệu được đưa vào trong lược đồ này.

- Lược đồ ngoài (External Schema) biểu diễn cho một người dùng/ ứng dụng cụ thể. Thông thường một lược đồ liên bang thường rất lớn, khó thay đổi, để giúp cho việc tùy biến trên từng đối tượng người dùng/ ứng dụng một cách linh hoạt bằng các mô hình dữ liệu khác nhau, hoặc giúp cho việc bổ sung tính toàn vẹn dữ liệu trên từng ứng dụng được dễ dàng.

Xử lý truy vấn trên FDBS: Khi một ứng dụng thực hiện truy vấn trên FDBS thì nảy sinh vấn đề là các đối tượng truy vấn có thể nằm hoàn toàn trên CSDL cục bộ, nhưng cũng có thể đối tượng đó không nằm hoàn toàn trên một CSDL cục bộ mà còn nằm trên các CSDL cục bộ khác. Một lược đồ thực hiện truy vấn trên FDBS được thể hiện như ở Hình 4.



Hình 4. Lược đồ xử lý truy vấn trong FDBS

Theo Hình 4, ta thấy khi nhập câu truy vấn dữ liệu toàn cục, hệ thống sẽ xem xét có thực hiện truy vấn từ các nguồn dữ liệu không cục bộ hay không. Nếu cần thực hiện truy vấn từ các nguồn dữ liệu không cục bộ, hệ thống sẽ kiểm tra nguồn dữ liệu thuộc CSDL quan hệ hay là phi CSDL quan hệ. Nếu nguồn dữ liệu thuộc CSDL quan hệ, hệ thống sẽ gọi thủ tục từ xa cho việc truy vấn CSDL quan hệ. Nếu nguồn dữ liệu là phi CSDL quan hệ thì hệ thống sẽ gửi câu truy vấn tới bộ xử lý truy vấn (Wrapper) để thực hiện truy vấn tùy thuộc vào loại dữ liệu này. Cuối hệ thống sẽ tổ hợp các kết quả truy vấn để được kết quả truy vấn toàn cục.

Vấn đề xử lý xung đột trong tích hợp lược đồ liên bang:

Các loại xung đột ở mức lược đồ:

- Xung đột tên: Hai khả năng xung đột liên quan đến tên trong lược đồ dữ liệu. Thứ nhất, các lược đồ có tên khác nhau nhưng có cùng một ngữ nghĩa. Cách giải quyết xung đột loại này là ánh xạ chúng tới cùng một tên chung. Khả năng thứ hai là các lược đồ có cùng một tên nhưng ngữ nghĩa lại khác nhau. Giải pháp khắc phục cho trường hợp này là thêm tiền tố trước tên của mỗi lược đồ xuất.

- Xung đột về cấu trúc: Cùng một ngữ nghĩa nhưng được biểu diễn bởi các cấu trúc dữ liệu khác nhau.

- Xung đột về ràng buộc: Hai kiểu ràng buộc toàn vẹn dữ liệu cần quan tâm nghiên cứu trong FDBS là ràng buộc toàn vẹn tính độc lập của ứng dụng và ràng buộc toàn vẹn dữ liệu của ứng dụng. Ràng buộc toàn vẹn tính độc lập của ứng dụng bao gồm ràng buộc về khóa, quy tắc ràng buộc thực thể, ràng buộc toàn vẹn tham chiếu và các phụ thuộc hàm liên quan đến quá trình chuẩn hóa.

Các loại xung đột ở mức dữ liệu:

- Xung đột đặt tên.

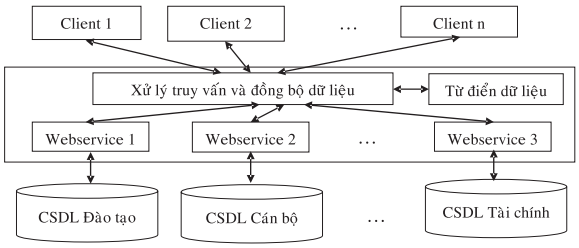
- Xung đột về biểu diễn dữ liệu: Ngữ nghĩa dữ liệu có thể được hiểu khác nhau trong các mô hình biểu diễn khác nhau [1]. Chẳng hạn, thực thể địa chỉ có thể là một quốc gia trong mô hình này, nhưng có thể là số nhà, tên đường, tên phường...

3. Vận dụng mô hình FDB thiết kế hệ thống tích hợp dữ liệu ở Trường Đại học Đồng Tháp

3.1. Thực trạng hệ thống phần mềm quản lý ở Trường Đại học Đồng Tháp

Trải qua nhiều giai đoạn phát triển, Trường Đại học Đồng Tháp hiện đang sử dụng nhiều phần mềm quản lý hoàn toàn độc lập với nhau như phần mềm quản lý đào tạo, phần mềm quản lý cán bộ, viên chức, phần mềm quản lý tài chính, phần mềm quản lý tài sản, phần mềm quản lý thư viện... Sự tồn tại của nhiều phần mềm trong cùng một tổ chức dẫn đến những bất cập: không có sự thống nhất thông tin giữa các phần mềm, khó có được thông tin mang tính tổng hợp cao từ các nguồn CSDL của các phần mềm, do đó ảnh hưởng đến công tác lãnh đạo, quản lý của tổ chức.

3.2. Đề xuất mô hình tích hợp dữ liệu ở Trường Đại học Đồng Tháp



Hình 5. Mô hình tích hợp dữ liệu ở Trường Đại học Đồng Tháp

Trong mô hình ở Hình 5, mô đun đồng bộ dữ liệu có nhiệm vụ kiểm tra dữ liệu từ các nguồn dữ liệu có ngữ nghĩa tương tự nhau, nếu có sự không thống nhất dữ liệu hệ thống có thể yêu cầu thực hiện cập nhật lại dữ liệu để đảm bảo sự đồng nhất dữ liệu. Để so sánh ngữ nghĩa giữa các nguồn dữ liệu, hệ thống sử dụng từ điển dữ liệu để định nghĩa sự tương đương giữa các thuộc tính dữ liệu.

Hiện nay, CSDL liên bang đang ở giai đoạn nghiên cứu lý thuyết, chưa có hệ quản trị CSDL liên bang hiệu quả mang tính thương mại. Vì vậy, ở Trường Đại học Đồng Tháp chúng tôi đề xuất mô hình tích hợp dữ liệu dựa trên nền tảng Web service ở Hình 5.

4. Kết luận

Bài báo đã trình bày những vấn đề cơ bản về tích hợp dữ liệu, đề xuất mô hình tổng quát và mô hình tích hợp dựa trên tiếp cận CSDL liên bang, nêu những đặc điểm cần quan tâm khi tích hợp hệ thống, các xung đột có thể xảy ra ở mức lược đồ, mức dữ liệu. Bài báo cũng đề xuất một mô hình tích hợp và cài đặt thử nghiệm hệ thống tích hợp ở Trường Đại học Đồng Tháp căn cứ vào các mô hình tích hợp dữ liệu đã trình bày và thông qua việc khảo sát các hệ phần mềm đang sử dụng.

Trong tương lai, chúng tôi sẽ đi sâu nghiên cứu việc tích hợp các lược đồ dựa trên nền tảng XML biểu diễn lược đồ liên bang, từ đó có thể thực hiện các truy vấn toàn cục một cách tổng quát./.

Tài liệu tham khảo

[1]. Deborah, L.Goldsmith, "Maintaining integrity in distributed and heterogenorous in database systems", *http://www.utdallas.edu/~bxt043000/Publications/Conference-Papers*.

[2]. Iskandar Ishak, Naomie Salim (2006), Database Integration Approaches for Heterogeneous Biological Data Sources: An overview, *Proceedings of the Postgraduate Annual Research Seminar 2006*, Faculty of Computer Science and Information System University Technology of Malaysia.

[3]. Patricia RodrÌguez, Gianolli and John Mylopoulos (2001), "A Semantic Approach to XML-based Data Integration", *University of Toronto, 6 Kingís College Road, Toronto, Canada M5S 3H5 {prg,jm}@cs.toronto.edu*.

[4]. Tony Schaller, Omran A. Bukhres, Ahmed K. Elmagarmid (1993), "The Integration of Database Systems", *Computer Science Technical Reports,* p. 1061.

[5]. A. P. Seth, J. A. Larson (1990), "Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases", *ACM Computing Surveys,* 22 (3).

INTEGRATED MODEL BASED ON FEDERATED DATABASE FOR BUILDING INTEGRATION INFROMATION SYSTEM IN DONG THAP UNIVERSITY

Summary

Information system integration has been quite long interested in research, and currently there are many different types of software made by independent organisations at different times. Therefore, it calls for integrating different software types to build overall information for management, operational activities. This paper gives a brief overview of the approach to data integration, federated database (FDB) and integrated data model based on FDB. In particular, the paper proposes one integrated data model that can be installed at the University of Dong Thap.

Keywords: federated database, multidatabase, data integration.